Approximate inference techniques with expectation constraints

Tom Heskes
Computer Science
Radboud University Nijmegen
T.Heskes@science.ru.nl

Manfred Opper
ISIS, School of Electronics and Computer Science
University of Southampton
mo@ecs.soton.ac.uk

Wim Wiegerinck
SNN
Radboud University Nijmegen
W.Wiegerinck@science.ru.nl

Ole Winther
Informatics and Mathematical Modelling
Technical University of Denmark
owi@imm.dtu.dk

Onno Zoeter*
SNN
Radboud University Nijmegen
O.Zoeter@science.ru.nl

September 29, 2005

Abstract

This article discusses inference problems in probabilistic graphical models that often occur in a machine learning setting. In particular it presents a unified view of several recently proposed approximation schemes. Expectation consistent approximations and expectation propagation are both shown to be related to Bethe free energies with *weak consistency constraints*, i.e. free energies where local approximations are only required to agree on certain statistics instead of full marginals.

1 Introduction

Probabilistic graphical models, such as Bayesian networks, Markov random fields and Boltzmann machines, are powerful tools for machine learning and reasoning in domains with uncertainty. Unfortunately, exact inference in probabilistic models is in many interesting cases numerically intractable. It requires the computation of marginal distributions and normalizing constants either through high dimensional integrations or sums over a number of values which increases exponentially with the number of variables. Hence, there is a great demand for techniques which give sensible approximations in polynomial time.

The field of variational inference had been dominated for a while by the so–called mean field method, where the full high-dimensional distribution is approximated by the closest one in a

*Corresponding author

tractable family. While this is a fairly general concept giving a bound on free energies (the negative logarithm of the intractable normalizer of the distributions), the accuracy of the method in predicting marginal moments is for some applications (e.g. in decoding and other signal processing applications) not sufficient. More recently, a variety of techniques for probabilistic inference which directly seek to approximate these marginal moments, have been newly developed or rediscovered. Ideas like the Bethe free energy approximations (introduced into machine learning by Yedidia et al. (2001)), Thomas Minka's EP framework (Minka, 2001b), the EC approach from Opper and Winther (2005) (generalizing their adaptive TAP scheme) and the cavity method (Mézard et al., 1987) seem to provide widely applicable concepts for approximate inference.

Unfortunately (as a variety of discussions within the machine learning community have indicated) there is some confusion about the meaning, applicability and relations between these approximations. It can be frustrating even for experts to compare derivations which are written for different scientific communities using a variety of different notations. Hence the goal of this paper is to address some of these issues, providing short derivations and cross connections (as we understand them) for the above mentioned approaches using a coherent notation. We emphasize that this is not intended to be a review article. Readers interested in complementary views of variational inference techniques we refer to (Wainwright and Jordan, 2003) and (Ikeda et al., 2000). We believe that by introducing a unification we can give some alternative points of view about the approximations which may enlarge their applicability to problems for which they had not been originally designed.

The paper is organized as follows. In Section 2 we define the main inference problems and introduce notation. It defines the *factor graph* as a framework for representing models and for representing choices in approximation schemes and introduces the *sum-product* for exact inference on trees or approximate inference on graphs with cycles. Section 3 reviews expectation propagation (EP) and expectation consistent (EC) approximations. Expectation propagation iteratively improves approximations by projecting onto tractable distributions. Expectation consistency approximations derive from the cavity method in statistical physics. Section 4 unifies both approximation methods from Section 3 by starting from a Bethe free energy and introducing the crucial concept of *weak consistency constraints*.

2 Probabilistic inference

2.1 Computing partition sums and marginal distributions

Probabilistic inference is the problem of computing the posterior probabilities of unobserved model variables $X = \{X_1, \ldots, X_N\}$ given the observations D of other model variables. The posterior probability $P(X = x|D)$, where X denotes the stochastic variable and x a particular realization, can be used in many ways e.g. to make forecasts about future data values. These are then expressed as certain expectations (averages) with respect to the posterior distribution. The goal of this paper is to address one of the key technical problems of this conceptually simple approach which lies in the practical computation of such expectations.

Our starting point is some probability distribution $p(x)$ which is assumed to be defined in terms of a given potential[1] $\Psi(x)$ and an unknown normalization Z,

$$p(x) = \frac{\Psi(x)}{Z} \,. \tag{1}$$

This structure is immediately present in the posterior distribution discussed above, with[2] $p(x) \equiv P(X = x|D)$, $\Psi(x) \equiv P(X = x, D)$ and normalization $Z \equiv P(D) = \sum_{X=x} P(X = x, D)$. It is also encountered in statistical models from physics, such as the Ising model with spin variables

[1] In physics, often the representation $\Psi(x) = \exp(-\psi(x))$ is used, in which $\psi(x)$ is called a potential.

[2] Here we focus for the moment on discrete variables, our main interest in this paper. Similar definitions apply to continuous variables.

$x_i = \pm 1$,

$$p(x) = \frac{1}{Z} \exp\left\{ \sum_{i,j} J_{ij} x_i x_j + \sum_i \theta_i x_i \right\} , \tag{2}$$

again with the partition sum Z such that the probability distribution normalizes to 1. The problem of inference is already apparent when one tries to compute the normalizer of the distribution which typically requires the computation of an in N exponentially large sum or an intractable integral (for continuous x):

$$Z = \sum_x \Psi(x) .$$

In statistical physics, Z is called the partition sum or function and $-\log Z$ is the corresponding free energy. A similar inference problem is the computation of a marginal distribution of a subset of variables $x_C \subset x$ which involves an exponential sum (or integral) over the variables outside the subset of interest,

$$p(x_C) = \sum_{x_{\setminus C}} p(x) .$$

Approximate computations of Z and marginal distributions $p(x_i)$ from which other expectations can be derived, will be the central topic of the paper.

We spend the remainder of this subsection to some notation. Expectations of a function $h(x)$ of the random variables x over a distribution p will be denoted by $\langle\, h(x) \,\rangle$, where we will add an index such as $\langle\, h(x) \,\rangle_p$, when it is not evident which distribution is used. We will use both sum and integral notation, depending on whether x is best viewed as a discrete or continuous variable. We stress that many equations with sum notation for discrete variables directly transfer to the same equation with the sum replaced by an integral for continuous variables, and viceversa.

2.2 Factor graphs

In many practical applications the model potential is expressed as a product of *factors*, (or sometimes called cluster potentials) labeled by α,

$$\Psi(x) = \prod_\alpha \Psi_\alpha(x_\alpha) \tag{3}$$

in which x_α denotes the variable vector restricted to the domain of Ψ_α.

This factorization can graphically be represented by a bi-partite graph known as a *factor graph* (Kschischang et al., 2001). In the factor graph, factors are represented by rectangles and variables are denoted by ovals. Each variable node x_i is connected by an undirected link to every factor node $\Psi_\alpha(x_\alpha)$ that contains the variable in their domain, $x_i \in x_\alpha$.

When two variables x_i and x_j always occur together in a factor, the two can be grouped together into a single (clustered) variable x_β which has more states. In the remainder of the paper we will use the notation x_β for the (clustered) variable nodes in the factor graphs, thereby including the factor graphs with the original variables as nodes as a special case. In other words, the convention in this paper is to label factor nodes by α and variable nodes by β. Since there is such a free choice in defining variables and factors, we will occasionally refer to x_α and x_β as *clusters*. In this terminology factor nodes are often called *outer clusters* and variable nodes *inner clusters*.

From (3) it is obvious that the factor nodes should span the whole domain:

$$\bigcup_\alpha x_\alpha = x .$$

For each variable node β there should be at least one factor node α that fully subsumes it:

$$\forall_\beta \exists_\alpha \text{ such that } x_\alpha \cap x_\beta = x_\beta .$$

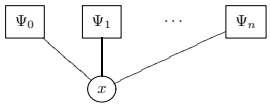

Figure 1: A factor graph corresponding to the i.i.d. assumption in Bayesian statistics. See Example 1.

Last but not least, in this paper we restrict ourselves to the case of non-overlapping variable nodes, i.e.,

$$x_\beta \cap x_{\beta'} = \emptyset \quad \forall_{\beta \neq \beta'} .$$

The more general case of overlapping variable nodes can be handled with so-called region graphs and leads to Kikuchi-based approximations (Yedidia et al., 2001).

The neighbor sets of α and β in the factor graph are denoted by N_α and N_β respectively. These sets have cardinalities $n_\alpha \equiv |N_\alpha|$ and $n_\beta \equiv |N_\beta|$.

The mapping from the original model to a factor graph is not unique. Two factors $\Psi_{\alpha'}$ and $\Psi_{\alpha''}$ can always be combined by taking their product,

$$\Psi_\alpha(x_\alpha) \equiv \Psi_{\alpha'}(x_{\alpha'})\Psi_{\alpha''}(x_{\alpha''}) ,$$

where $x_\alpha = x_{\alpha'} \cup x_{\alpha''}$. In addition, the factors are not unique. If two factors are connected via a variable node x_β we can shift these factors by any non-zero factor $\Phi_\beta(x_\beta)$:

$$\Psi'_\alpha(x_\alpha) = \Psi_\alpha(x_\alpha)\Phi_\beta(x_\beta), \quad \Psi'_{\alpha'}(x_{\alpha'}) = \frac{\Psi'_{\alpha'}(x_{\alpha'})}{\Phi_\beta(x_\beta)} .$$

Finally, we note that two different factors may have the same domain, $x_\alpha = x_{\alpha'}$. This will turn out useful in the discussion of approximate inference methods.

Example 1 *Let us consider the following problem, often encountered in Bayesian statistics and machine learning. We have a joint probability model $P(X,Y) = P(Y|X)P(X)$, with $P(X)$ the prior distribution and $P(Y|X)$ a generative model. We observe a data set $D = \{y_1, \ldots, y_i, \ldots, y_n\}$ with different realizations of the random variable Y and would like to derive (properties of) the posterior distribution $P(X = x|D)$. Assuming the data points y_i independently and identically distributed (i.i.d.), Bayes' rule yields*

$$P(X = x|D) = \frac{P(X = x)\prod_i P(Y = y_i|X = x)}{P(D)} , \tag{4}$$

with $P(D) = \sum_{X=x} P(X = x)\prod_i P(Y = y_i|X = x)$ the proper normalization. This can be written in the form (3) e.g. with definitions $\Psi_0(x) = P(X = x)$ and $\Psi_i(x) = P(X = x|Y = y_i)$. The corresponding factor graph is visualized in Figure 1. There are $n + 1$ factor nodes and 1 variable node, linked to all factor nodes. All factor nodes have the same, namely the complete, domain.

Example 2 *The two-dimensional Ising lattice of $K \times L$ nodes is (without external fields θ_i)*

$$p(x) = \frac{1}{Z}\exp\left\{\sum_{i=1}^{K-1}\sum_{j=1}^{L} J_{i,j;i+1,j}x_{i,j}x_{i+1,j} + \sum_{i=1}^{K}\sum_{j=1}^{L-1} J_{i,j;i,j+1}x_{i,j}x_{i,j+1}\right\} .$$

With the definitions

$$\begin{aligned}\Psi_{i,j;i+1,j}(x_{i,j}, x_{i+1,j}) &= \mathrm{e}^{x_{i,j}J_{i,j;i+1,j}x_{i+1,j}} \quad \textit{and} \\ \Psi_{i,j;i,j+1}(x_{i,j}, x_{i,j+1}) &= \mathrm{e}^{x_{i,j}J_{i,j;i,j+1}x_{i,j+1}}\end{aligned}$$

we obtain the factor graph from Figure 2 (left). This choice results in a model with loops. By grouping variables together in vertical chains $x'_j \equiv [x_{1,j}; x_{2,j}; \ldots; x_{K,j}]$, and combining factors accordingly we obtain the factor graph from Figure 2 (right).

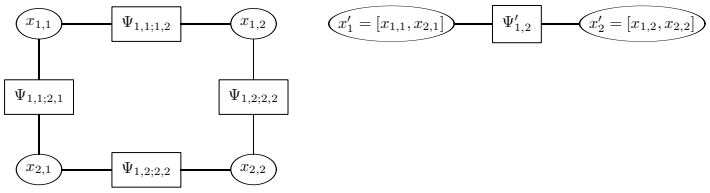

Figure 2: The left graph shows a straightforward factor graph representation of a two-dimensional Ising lattice. The right graph shows a representation of the same model. In this representation variables and factors are grouped together such that the resulting graph is a tree. See Example 2.

2.3 The sum-product algorithm

2.3.1 Trees

A factor graph is called a tree if there is at most one path from a node in the graph to another node in the graph. In a tree, the joint probability can be expressed in terms of its *cluster marginals* $p(x_\alpha)$ and $p(x_\beta)$,

$$p(x) = \frac{\prod_\alpha p(x_\alpha)}{\prod_\beta p(x_\beta)^{n_\beta - 1}} \ , \tag{5}$$

where

$$p(x_\alpha) = \sum_{x_{\setminus \alpha}} p(x) \text{ and } p(x_\beta) = \sum_{x_{\setminus x_\beta}} p(x) \ .$$

The other way around, any set of cluster marginals $\{p_\alpha(x_\alpha), p_\beta(x_\beta)\}$ that satisfies the following conditions

$$p_\alpha(x_\alpha) \geq 0 \text{ and } p_\beta(x_\beta) \geq 0 \quad \forall \alpha, \beta \ \text{ (non-negativity)} \tag{6}$$

$$\sum_{x_\alpha} p_\alpha(x_\alpha) = 1 \text{ and } \sum_{x_\beta} p_\beta(x_\beta) = 1 \quad \forall_{\alpha,\beta} \ \text{ (normalization)} \tag{7}$$

$$p_\alpha(x_\beta) \equiv \sum_{x_{\alpha \setminus \beta}} p_\alpha(x_\alpha) = p_\beta(x_\beta) \quad \forall_{\beta, \alpha \in N_\beta} \ \text{ (local consistency)} \tag{8}$$

defines a global distribution on a tree as in (5), with marginals

$$p(x_\alpha) = p_\alpha(x_\alpha) \quad \text{and} \quad p(x_\beta) = p_\beta(x_\beta) \ .$$

2.3.2 The discrete case

We can view (5) as an alternative factorization to (3). It can be found by applying a message-passing algorithm called belief propagation or, in the context of factor graphs, the sum-product algorithm (Kschischang et al., 2001; Pearl, 1988). For multinomial models with just discrete variables, the messages $M_{\beta \to \alpha}(x_\beta)$ are initially all set to 1, and updated as

$$\begin{aligned} M_{\alpha \to \beta}(x_\beta) &= \sum_{x_{\alpha \setminus \beta}} \Psi_\alpha(x_\alpha) \prod_{\beta' \in N_\alpha \setminus \beta} M_{\beta' \to \alpha}(x_{\beta'}) \ , \\ M_{\beta \to \alpha}(x_\beta) &= \prod_{\alpha' \in N_\beta \setminus \alpha} M_{\alpha' \to \beta}(x_\beta) \ . \end{aligned} \tag{9}$$

After convergence of the procedure, which can be guaranteed within a finite number of updates if no edges are neglected, the marginals follow from

$$\begin{aligned} p(x_\alpha) &\propto \prod_{\beta \in N_\alpha} M_{\beta \to \alpha}(x_\beta) \Psi_\alpha(x_\alpha) \ , \\ p(x_\beta) &\propto \prod_{\alpha \in N_\beta} M_{\alpha \to \beta}(x_\beta) \ . \end{aligned} \tag{10}$$

Furthermore, the partition function reads

$$Z = \sum_{x_\beta} \prod_{\alpha \in N_\beta} M_{\alpha \to \beta}(x_\beta) \quad \text{for any } \beta \; .$$

The computation time is linear in the number of clusters and exponential in the cluster size. For example, the computation time in the $K \times L$ Ising grid of example 2 in its tree representation is linear in K and exponential in L.

2.3.3 Generalization to the exponential family

Message passing in trees with discrete variables can be interpreted as a special case of message passing in trees in the exponential family. In these models the cluster potentials are parameterized as

$$\Psi_\alpha(x_\alpha) = \exp\left\{\kappa_\alpha^\top \phi_\alpha(x_\alpha)\right\} \; . \tag{11}$$

The vector of functions $\phi_\alpha(x_\alpha)$ is the so-called sufficient statistics of α. The sufficient statistics are fixed, and as such they are part of the definition of the model class. The parameters in (11) are given by κ_α, the parameter vector of cluster α. The multinomial model discussed above can be written in exponential form by defining the sufficient statistics to be a vector with a component for each state x'_α:

$$\begin{aligned} \phi_\alpha^{x'_\alpha}(x_\alpha) &= 1 \text{ if } x_\alpha = x'_\alpha \\ \phi_\alpha^{x'_\alpha}(x_\alpha) &= 0 \text{ otherwise} \end{aligned}$$

The messages at all times remain within this exponential family if the potentials have a parametrization that is preserved under marginalization from outer clusters to inner clusters. This condition is valid if for all $\alpha \in N_\beta$, and any parameter vector κ_α the marginal of the outer cluster can be written as a potential on the inner cluster β, parameterized in an exponential family form,

$$\int dx_{\alpha \setminus \beta} \; \exp\left\{\kappa_\alpha^\top \phi_\alpha(x_\alpha)\right\} = \exp\left\{\kappa_\beta^\top \phi_\beta(x_\beta)\right\}$$

in which κ_β is a parameter vector of inner cluster β and $\phi_\beta(x_\beta)$ its statistics. We will refer to this condition as "closed under marginalization". If the exponential family is indeed closed under marginalization, the messages in (9) are easily shown to remain of exponential form and can be fully expressed in terms of their parameters:

$$\begin{aligned} M_{\alpha \to \beta}(x_\beta) &= \exp\left\{\mu_{\alpha \to \beta}^\top \phi_\beta(x_\beta)\right\} \\ M_{\beta \to \alpha}(x_\beta) &= \exp\left\{\mu_{\beta \to \alpha}^\top \phi_\beta(x_\beta)\right\} \; . \end{aligned}$$

Two well-known families that are closed under marginalization and hence allow for efficient exact inference algorithms for tree structured models are models with multinomial potentials and models with Gaussian potentials. The complexity of the message passing algorithms in such models is linear in the number of clusters times the complexity of cluster marginalization. In discrete models, this latter operation is exponential in the cluster size. In Gaussian models, marginalization in a cluster of N nodes involves the inversion of an $N \times N$ matrix, which is $\mathcal{O}(N^3)$.

2.3.4 Loopy belief propagation

The sum-product algorithm (9) gives the exact marginals in trees. As an iteration scheme, however, it can be implemented in non-trees as well. Pearl (1988) proposed to apply this scheme to non-trees as an approximation method for computing cluster marginals. This approximation algorithm is called *loopy belief propagation*. Loopy belief propagation is applicable to multinomial models, as well as models in the exponential family with potentials that are preserved under marginalization

as explained in the previous section, e.g, models with Gaussian potentials (Weiss and Freeman, 2001). Until recently, a disadvantage of the method was its heuristic character, and the absence of a convergence guarantee. Often, the algorithm gives surprisingly good solutions, but sometimes the algorithm fails to converge (Murphy et al., 1999).

Yedidia et al. (2001) showed that the fixed points of loopy belief propagation are actually stationary points of the Bethe free energy from statistical physics (see also Heskes (2003)). This can be considered as a breakthrough in the field. The Bethe free energy provides a firm theoretical basis of loopy belief propagation and it served as a basis for more advanced methods, such as generalized belief propagation (Yedidia et al., 2001). Furthermore, it provides a unifying framework relating loopy belief propagation to several other approximation methods such as mean field theory (Jordan et al., 1999), and last but not least, it provides a framework to solve the convergence problem by the existence of an objective function which can be minimized directly (Welling and Teh, 2001; Yuille, 2002; Teh and Welling, 2002; Heskes et al., 2003).

3 Two approximate inference methods using expectations

This section reviews expectation propagation (Minka, 2001b) and expectation consistent approximations (Opper and Winther, 2005). Expectation propagation has been proposed in the machine learning community as an extension to assumed density filtering. Expectation consistency can be viewed as a generalization of the ADATAP (Opper and Winther, 2001b,a) approximation introduced in the physics community.

3.1 Expectation propagation

3.1.1 Assumed density filtering

In (Minka, 2001b), expectation propagation (EP) is motivated as an iterative refinement of assumed density filtering (ADF). Assumed density filtering aims to approximate the posterior $P(X = x|D)$ from (4) that arises in the context of Bayesian machine learning, example 1. This approximation, call it $\tilde{p}(x)$, is chosen to be a certain (convenient) parametric distribution in the exponential family, specified by the vector $\phi(x)$ of sufficient statistics.

The approximation is initialized by the prior $\tilde{p}(x) = P(X = x) = \Psi_0(x)$, which is here assumed to be part of this family. Then each data point y_i is visited exactly once in a sequential way. For each data point, we first incorporate the data point by multiplying in the corresponding potential:

$$q_i(x) \propto \tilde{p}(x)\Psi_i(x) \; .$$

If $\Psi_i(x)$ is not within the exponential family, neither is $q_i(x)$. Therefore, the next step is to project $q_i(x)$ back to the exponential family by minimization of the Kullback-Leibler divergence

$$\mathrm{KL}(q_i(x)||\tilde{p}^{\mathrm{new}}(x)) \equiv \int dx \; q_i(x) \log \frac{q_i(x)}{\tilde{p}^{\mathrm{new}}(x)} \; ,$$

under the constraint that the new approximation $\tilde{p}^{\mathrm{new}}(x)$ is again in the family. The minimization under these constraints is equivalent to moment matching, i.e. the minimizing distribution $\tilde{p}^{\mathrm{new}}(x)$ is the unique distribution that satisfies $\langle \phi(x) \rangle_{q_i} = \langle \phi(x) \rangle_{\tilde{p}^{\mathrm{new}}}$. After updating $\tilde{p}(x) = \tilde{p}^{\mathrm{new}}(x)$, the procedure continues by incorporating the next data point.

3.1.2 Refining term effects

The outcome of assumed density filtering typically depends on the order in which the data points are incorporated. Expectation propagation can be viewed as a refinement of the assumed density filtering approximation that tries to compensate for this ordering anomaly. It does so by keeping track of and refining the approximate contributions or *term effects* $M_i(x)$ of each data point y_i to the approximate posterior $\tilde{p}(x)$.

Initially, all term effects $M_i(x)$ are set to 1 and as before $\tilde{p}(x)$ is initialized to the prior $\Psi_0(x)$. To refine a term effect, we first *take it out* by dividing through it:

$$\tilde{p}_{\setminus i}(x) \propto \frac{\tilde{p}(x)}{M_i(x)} \, .$$

Next, we create a new approximation by *putting back* the exact contribution $\Psi_i(x)$, multiplying it in:

$$q_i(x) \propto \tilde{p}_{\setminus i}(x)\Psi_i(x) \, .$$

The distribution $q_i(x)$ is typically outside the chosen family. Therefore, as with assumed density filtering, we project it back to this family yielding the new $\tilde{p}^{\text{new}}(x)$. The refinement is updated as the new approximate posterior (after incorporation of y_i) divided by the one without the term effect:

$$M_i^{\text{new}}(x) \propto \frac{\tilde{p}^{\text{new}}(x)}{\tilde{p}_{\setminus i}(x)} \propto \frac{\tilde{p}^{\text{new}}(x) M_i(x)}{\tilde{p}(x)} \, . \tag{12}$$

It is then easy to see that when we start out with $M_i(x)$ having the particular exponential form (which we do when we initialize them to 1), it will always have this form. After updating $\tilde{p}(x) = \tilde{p}^{\text{new}}(x)$ and $M_i(x) = M_i^{\text{new}}(x)$, the procedure continues by refining the next term effect.

Once "refining" all term effects sequentially is equivalent to assumed density filtering. Expectation propagation typically iterates in random order until convergence (which is not guaranteed).

3.1.3 The general case

The above example is special in the sense that all potentials are functions over the complete domain x. The corresponding factor graph, Figure 1, then contains just a single variable node. Expectation propagation can handle the more general case of localized potentials as well.

In its simplest version, expectation propagation takes the approximating distribution fully factorized, as a product of distributions of non-overlapping inner clusters β,

$$\tilde{p}(x) = \prod_{\beta} \tilde{p}_\beta(x_\beta) \, . \tag{13}$$

Furthermore, each distribution $\tilde{p}_\beta(x_\beta)$ is chosen to be in a convenient exponential family defined by a vector $\phi_\beta(x_\beta)$ of sufficient statistics.

By definition, the approximating distribution can also be written as a product of term effects,

$$\tilde{p}(x) \propto \prod_{\alpha} M_\alpha(x_\alpha) \, , \tag{14}$$

where $M_\alpha(x_\alpha)$ corresponds to the contribution of the potential $\Psi_\alpha(x_\alpha)$. For (13) and (14) to be consistent, the term effects should factorize over β as well and we can write

$$M_\alpha(x_\alpha) = \prod_{\beta \in N_\alpha} M_{\alpha\to\beta}(x_\beta) \, , \tag{15}$$

where, as will become clear later on, we can interpret the terms $M_{\alpha\to\beta}(x_\beta)$ as messages. Reshuffling, we can then also express $\tilde{p}_\beta(x_\beta)$ in terms of these messages:

$$\tilde{p}_\beta(x_\beta) \propto \prod_{\alpha \in N_\beta} M_{\alpha\to\beta}(x_\beta) \, . \tag{16}$$

And finally, with $\tilde{p}_\beta(x_\beta)$ of a particular exponential form, the messages will have the same form:

$$M_{\alpha\to\beta}(x_\beta) = \exp\left\{\mu_{\alpha\to\beta}^\top \phi_\beta(x_\beta)\right\}$$

parameterized by the vector $\mu_{\alpha\to\beta}$.

Expectation propagation can be initialized by random parameter vectors $\mu_{\alpha\to\beta}$ for the effects (such that the sums $\sum_{\alpha\in N_\beta}\mu_{\alpha\to\beta}$ yield valid parameter vectors for the distributions $p_\beta(x_\beta)$). Then the term effects are iteratively refined. The refinement of the term effects of α is carried out as follows. First, we take out the term effect $M_\alpha(x_\alpha)$ and put back the exact potential $\Psi_\alpha(x_\alpha)$ yielding

$$q_\alpha(x) = \frac{\tilde{p}(x)}{M_\alpha(x_\alpha)}\Psi_\alpha = q_\alpha(x_\alpha)\prod_{\beta\notin N_\alpha} p_\beta(x_\beta)\,,$$

with

$$q_\alpha(x_\alpha) \propto \prod_{\beta\in N_\alpha}\prod_{\alpha'\in N_\beta\setminus\alpha} M_{\alpha'\to\beta}(x_\beta)\Psi_\alpha(x_\alpha)\,,$$

which is easily derived from (**??**) through (16). Next, we project q_α back onto the factorized exponential family by minimizing $\mathrm{KL}(q_\alpha||\tilde{p}^{\mathrm{new}})$. This yields for each $\beta\in N_\alpha$ a new moment-matched exponential distribution

$$\tilde{p}^{\mathrm{new}}_\beta(x_\beta) = \exp(\gamma_\beta^\top\phi_\beta(x_\beta)) \tag{17}$$

with γ_β such that

$$\langle\phi_\beta(x_\beta)\rangle_{\tilde{p}^{\mathrm{new}}_\beta} = \langle\phi_\beta(x_\beta)\rangle_{q_\alpha}\ .$$

The other $\beta\notin N_\alpha$ are not affected by the refinement of α. Finally, the new term effects of α are found similarly to (12). In terms of the parameters they are

$$\mu^{\mathrm{new}}_{\alpha\to\beta}(x_\beta) = \gamma_\beta - \sum_{\alpha'\in N_\beta\setminus\alpha}\mu_{\alpha'\to\beta}(x_\beta)\,. \tag{18}$$

This procedure of refinements is again iterated in random order until convergence (not guaranteed).

Expectation propagation applied to trees in an exponential family that is closed under marginalization reduces to the sum-product algorithm from Section 2.3 and gives the exact marginals on inner and outer clusters.

It can be shown that the term effects in expectation propagation exactly correspond to the messages in the sum-product algorithm (Minka, 2001b). Thus, expectation propagation can be motivated from a tree-like argument, with in addition the assumption that the inner cluster marginals are approximately distributions from the chosen exponential family (e.g. Gaussians). In a multinomial model, expectation propagation reduces to loopy belief propagation.

3.2 Expectation consistent (EC) approximations

In the following we will discuss the *expectation consistent* (EC) approximation introduced recently by (Opper and Winther, 2005). It is a generalization of the ADATAP approximation (Opper and Winther, 2001b,a), which itself is motitaved by the cavity method.

The cavity method (Mézard et al., 1987) can be viewed as a technique for deriving a closed set of equations for approximate marginal distributions of probabilistic models. These equations are often referred to as TAP equations (named after the physicists Thouless, Anderson & Palmer) (Thouless et al., 1977). The method has its origin in the statistical physics of disordered magnets with infinitely ranged random interactions. For its application to problems in the area of machine learning the method – so far in its simplest version of a single pure state – had to be tuned to allow for more complex probabilistic models with structured (non-random) interactions (Opper and Winther, 2001b,a). The ideas presented in these papers were inspired by earlier work of Parisi and Potters (Parisi and Potters, 1995). For similar work on TAP equations for Ising models, see (Kappen and Rodríguez, 1999).

We will next try to motivate the EC approximation for models with pairwise interactions such as the Ising model (2). We will use cavity ideas in a fairly informal way, refrain from giving formal definitions of "cavity fields" etc.

The partition function of such models can be written in the form

$$Z = \int dx\ \Psi_1(x)\Psi_2(x)\,,$$

with the factors

$$\begin{aligned}\Psi_1(x) &= \prod_i \psi_i(x_i) \\ \Psi_2(x) &= \exp\left[x^T J x\right]\ .\end{aligned}$$

If the factor Ψ_2 would be absent, all spins would be noninteracting and the model could be trivially solved. To deal approximately for the neglected interactions we could introduce for each single variable a cavity field which accounts for the influence of all other variables that are connected to it in a mean field type of fashion. Doing this independently for each variable x_i would lead to an approximation for the partition function

$$Z \approx Z_1(\Lambda) = \int dx\ \Psi_1(x)\ \exp[\Lambda^T \phi(x)] \tag{19}$$

where we set $\phi(x) = (x_1, x_1^2, x_2, x_2^2, \ldots, x_N, x_N^2)$, (see (Opper and Winther, 2005) for other possibilities). The linear terms act as simple mean field terms, and the quadratic ones (which are trivial for Ising variables, but are needed for continuous x_i), can be understood by assuming a Gaussian statistics for the random field $h_i \equiv \sum_j J_{ij} x_j$ measured in the "cavity" which is created when variable x_i is removed from the system. Such a Gaussian assumption seems reasonable because h_i is composed of a sum of many weakly dependent variables. It can be perfectly justified (assuming a single ergodic state) for models with quenched random interactions in the"thermodynamic" limit $N \to \infty$. Nevertheless, even in this limit, the naive approximation (19) would be plain wrong because important effects of the interactions are still neglected. We will motivate a correction which also helps us to compute the vector of parameters Λ in a self-consistent way.

To do this we express the exact partition function Z using Z_1 and a correction term:

$$Z = Z_1(\Lambda_1)\,\langle\, \Psi_2(x)\ \exp[-\Lambda_1^T \phi(x)]\,\rangle_{\,\tilde{p}_1}\,, \tag{20}$$

where the average at the right is defined through the distribution

$$\tilde{p}_1(x) = \frac{1}{Z_1(\Lambda_1)} \Psi_1(x)\ \exp\left[\Lambda_1^T \phi(x)\right] \tag{21}$$

Of course, this average cannot be computed efficiently, but we will again invoke a "cavity" type of argument assuming that the replacement of the average over the many (hopefully weakly dependent) variables using the exact distribution $\tilde{p}(x)$ by the average over an effective factorizing Gaussian distribution

$$q(x;\hat{\Lambda}) = \frac{1}{\hat{Z}(\hat{\Lambda})} \exp\left[\hat{\Lambda}^T \phi(x)\right] \tag{22}$$

could give a good approximation. This yields an approximation to the partition function given by

$$Z \approx \frac{Z_1(\Lambda) Z_2(\hat{\Lambda} - \Lambda)}{\hat{Z}(\hat{\Lambda})} \tag{23}$$

with

$$Z_2(\Lambda) = \int dx\ \Psi_2(x)\ \exp[\Lambda^T \phi(x)] \tag{24}$$

The parameter $\hat{\Lambda}$ is defined through the requirement that (21) and its approximation (22) should have the same expected statistics (expectation consistency) $\langle\, \phi \,\rangle_{\,q} = \langle\, \phi \,\rangle_{\,\tilde{p}_1}$ Finally, the parameter Λ is determined from the observation that the exact relation (20) is independent of Λ. Hence, a

good idea should be to make the approximation (23) stationary with respect to variations of Λ. This condition then leads to a further set of expectation consistency relations $\langle \phi \rangle_q = \langle \phi \rangle_{\tilde{p}_2}$, where $\tilde{p}_2$ is defined as in (21), but replacing Ψ_1 by Ψ_2.

It is not hard to show that these assumptions can also be expressed by the stationarity of the following approximate EC free energy with respect to variations of two sets of variables Λ_1 and Λ_2:

$$-\log Z^{\mathrm{EC}}(\Lambda_1, \Lambda_2) = -\log Z_1(\Lambda_1) - \log Z_2(\Lambda_2) + \log \hat{Z}(\Lambda_1 + \Lambda_2) \ ,$$

Example 3 *For Ising variables with $\psi_i(x_i) = [\delta(x_i + 1) + \delta(x_i - 1)] \exp[\theta_i x_i]$, the partition functions $Z_1(\Lambda_1)$ and $Z_2(\Lambda_2)$ can be computed in polynomial time. In fact, Z_1 completely factorizes over the variables. Setting $\Lambda_1(i) = (\gamma_i, \lambda_i)$ we can write*

$$\begin{aligned} Z_1(\Lambda_1) &= \prod_i \int dx_i \psi_i(x_i) \exp\left[\gamma_i x_i + \lambda_i x_i^2\right] = \prod_i \sum_{x_i = \pm 1} \exp\left[(\gamma_i + \theta_i) x_i + \lambda_i x_i^2\right] \\ &= \prod_i \left[2 \cosh(\gamma_i + \theta_i) \mathrm{e}^{\lambda_i}\right] \ . \end{aligned} \tag{25}$$

Z_2 is nothing but the normalizer for a multivariate Gaussian integral:

$$\begin{aligned} Z_2(\Lambda_2) &= \int dx \ \exp\left[(\gamma + \theta)^T x + x^T (\mathrm{diag}(\lambda) + J) x\right] \\ &= \sqrt{\frac{(4\pi)^N}{\det(-(\mathrm{diag}(\lambda) + J))}} \exp\left[-(\gamma + \theta)^T (\mathrm{diag}(\lambda) + J)^{-1} (\gamma + \theta)\right] \ . \end{aligned} \tag{26}$$

Note that λ cannot be chosen freely, but has to be restricted to values that make $-(\mathrm{diag}(\lambda) + J)$ positive definite, see (Opper and Winther, 2005) for a discussion of how to deal with this in practice. Finally

$$\hat{Z}(\Lambda) = \prod_i \left\{ \int dx_i \exp\left[(\gamma_i + \theta_i) x_i + \lambda_i x_i^2\right] \right\} = \prod_i \left\{ \sqrt{\frac{4\pi}{-\lambda_i}} \exp\left[\frac{-(\gamma_i + \theta_i)^2}{\lambda_i}\right] \right\} \ .$$

and

$$q(x) = \frac{1}{\hat{Z}(\Lambda)} \prod_i \exp\left[(\gamma_i + \theta_i) x_i + \lambda_i x_i^2\right]$$

It is possible to generalize the EC approximation to models with the more general type of factorization

$$p(x) = \frac{1}{Z} \prod_{\alpha=1}^{n} \Psi_\alpha(x) \ ,$$

and the corresponding intractable partition function

$$Z = \int dx \ \prod_\alpha \Psi_\alpha(x) \ .$$

For this case the EC approximation is obtained by extremizing an EC free energy of the form

$$-\log Z^{\mathrm{EC}}(\Lambda_1, \ldots, \Lambda_n) = -\sum_\alpha \log Z_\alpha(\Lambda_\alpha) + (n-1) \log \hat{Z}(\frac{1}{n-1} \sum_\alpha \Lambda_\alpha) \ . \tag{27}$$

with respect to the parameters Λ_α, where the Z_α are defined similar to (19,24). A solution is to be found using numerical methods. Several solutions may exist and the expectation consistent framework by itself does not provide a criterion to choose an optimal solution.

4 Unifying approximations: weak consistency constraints in Bethe free energies

The motivations for expectation propagation and expectation consistency are quite different. In this section, we will show how both approaches can be derived from a Bethe free energy with weak consistency constraints. Our arguments are closely related to the ones used for showing the relationship between loopy belief propagation and the classical Bethe free energy with strong consistency constraints (Yedidia et al., 2001). The original free energy corresponding to an expectation propagation algorithm is from Minka (2001a). The relationship with the Bethe free energy and the notion of weak consistency constraints is from Heskes and Zoeter (2002).

4.1 The Bethe free energy with weak constraints

4.1.1 A variational objective

Our starting point is (again) the probability distribution (1) with the factorization (3). We first cast the (intractable) calculation of

$$\log Z = \log \int dx \prod_\alpha \Psi_\alpha(x_\alpha)$$

as an optimization problem and then proceed by approximating the optimization problem. We adhere to the notational convention in the physics literature and add a minus to obtain

$$\begin{aligned} -\log Z &= \min_{\tilde{p}} \left[-\log Z + \mathrm{KL}\left(\tilde{p}(x)||p(x)\right)\right] \\ &= \min_{\tilde{p}} \left[-\log Z + \int dx\, \tilde{p}(x) \log \frac{\tilde{p}(x)}{Z^{-1}\prod_\alpha \Psi_\alpha(x_\alpha)} \right] \\ &= \min_{\tilde{p}} \left[-\sum_\alpha \int dx_\alpha\, \tilde{p}(x_\alpha) \log \Psi_\alpha(x_\alpha) + \int dx\, \tilde{p}(x) \log \tilde{p}(x) \right] \\ &\equiv \min_{\tilde{p}} F(\tilde{p})\,, \end{aligned} \tag{28}$$

where the minimization is over all valid distributions over the domain of x, i.e. $\tilde{p}(x) \geq 0$ for all x and $\int dx\, \tilde{p}(x) = 1$. We will refer to $F(\tilde{p})$ as the variational free energy.

Note that with the choice of adding $\mathrm{KL}\left(\tilde{p}(x)||p(x)\right)$ as a slack term, the two occurrences of $\log Z$ cancel and the optimization problem does not involve the intractable log partition function any more. Since $\mathrm{KL}\left(\tilde{p}(x)||p(x)\right)$ is positive and equals zero only if $\tilde{p}(x) = p(x)$ (Gibbs inequality), exact minimization of the above variational problem results in the true $-\log Z$. But as mentioned above, we assume that this is intractable.

4.1.2 Trees

In general, the entropy term in the free energy involves a summation over exponentially many states. In trees, the entropy can be simplified considerably. In a tree with outer clusters α and inner clusters β, the joint distribution $\tilde{p}(x)$ is fully specified in terms of locally consistent cluster marginals $\tilde{p}_\alpha(x_\alpha)$. By substitution of the representation in terms of marginals into the free energy, we find that F can be written as

$$\begin{aligned} F(\tilde{p}) = &-\sum_\alpha \int dx_\alpha\, \tilde{p}_\alpha(x_\alpha) \log \Psi_\alpha(x_\alpha) \\ &+ \sum_\alpha \int dx_\alpha\, \tilde{p}_\alpha(x_\alpha) \log \tilde{p}_\alpha(x_\alpha) - \sum_\beta (n_\beta - 1) \int dx_\beta\, \tilde{p}_\beta(x_\beta) \log \tilde{p}_\beta(x_\beta) \\ = &\; F^{\mathrm{B}}(\{\tilde{p_\alpha}, \tilde{p_\beta}\}) \end{aligned} \tag{29}$$

which is to be minimized under the consistency constraints (6), (7), and (8).

4.1.3 Bethe-type approximations

If the model does not allow for a tree factorization with reasonably sized clusters we can approximate F. (Structured) mean field approximations are obtained by restricting the set over which $\tilde{p}(x)$ is minimized (Saul et al., 1996; Wiegerinck, 2000). For instance, (28) can be minimized under the additional constraints that $\tilde{p}(x)$ is fully factorized

$$\tilde{p}(x) = \prod_n \tilde{p}_n(x_n) \,.$$

This approximating joint distribution ignores the existence of clusters α that are explicitly present in the cluster representation of the model. This is in contrast to the Bethe approximation, which takes by construction all the clusters into account. The Bethe approximation considers a set of locally consistent cluster marginals $\{\tilde{p}_\alpha, \tilde{p}_\beta\}$ rather than a restricted global joint distribution $\tilde{p}(x)$ as in mean field. In addition – despite the fact that the factor graph is in general not a tree – it makes the "tree-like" assumption for the free energy,

$$F(\tilde{p}) \approx F^{\mathrm{B}}(\{\tilde{p}_\alpha, \tilde{p}_\beta\}) \tag{30}$$

with F^{B} as defined in (29), which is again to be minimized under the constraints that the conditions (6), (7), and (8) hold.

If the original model contains loops, the tree-like form (5) with $\tilde{p}_\alpha(x_\alpha)$ for $p(x_\alpha)$ and $\tilde{p}_\beta(x_\beta)$ for $p(x_\beta)$, need not be normalized. Also, computing marginals over x_α and x_β from the product in (5), even after a possible normalization, does not in general retrieve $\tilde{p}_\alpha(x_\alpha)$ and $\tilde{p}_\beta(x_\beta)$. Hence they are sometimes referred to as "pseudo-marginals", see Wainwright and Jordan (2003) for a detailed discussion.

Since the negative entropy term is not derived from a global distribution there is no guarantee that minimizing (29) leads to a bound of $-\log Z$ as in the mean-field case.

Note also that, since we have restricted ourselves to factor graphs there are no two inner clusters β and β' such that $x_\beta \subset x_{\beta'}$. A generalization to approximations where overlaps overlap themselves is known as the Kikuchi free energy. We refer to Yedidia et al. (2001) for a more detailed discussion.

4.1.4 Weak consistency constraints

To make the connection with expectation propagation and expectation consistent approximation, we introduce the concept of weak constraints. First of all, instead of allowing any distribution $\tilde{p}_\beta(x_\beta)$, we constrain these to be part of a particular exponential family, characterized through the sufficient statistics $\phi_\beta(x_\beta)$. Next we relax the constraints (8) by requiring only consistency with respect to these sufficient statistics:

$$\langle \phi_\beta(x_\beta) \rangle_{\tilde{p}_\alpha(x_\beta)} = \langle \phi_\beta(x_\beta) \rangle_{\tilde{p}_\beta(x_\beta)} \,, \quad \forall_\beta \forall_{\alpha \in N_\beta} \,. \tag{31}$$

In Lauritzen (1992) the exponential family belief

$$r(x_\beta) \propto \mathrm{e}^{\gamma^\top \phi_\beta(x_\beta)} \quad \text{with } \gamma \text{ such that } \langle \phi_\beta(x_\beta) \rangle_q = \langle \phi_\beta(x_\beta) \rangle_{\tilde{p}_\alpha}$$

is called the *weak marginal* of $\tilde{p}_\alpha(x_\alpha)$. In words we then can say that (31) only requires the consistency of weak instead of strong marginals. To distinguish the local consistency constraint (8) from the concept of *weak consistency* (31) introduced above, we will refer to (8) as *strong consistency*.

Note that we do not enforce a particular parametric form of $\tilde{p}_\alpha(x_\alpha)$. However, at a minimum of the approximate free energy its form depends on the factors $\Psi_\alpha(x_\alpha)$ of the original model and the choice of sufficient statistics $\phi_\beta(x_\beta)$. The exact relationship is discussed in Section 4.2.

4.2 Finding stationary points of the free energy

4.2.1 The Lagrangian

In principle, we would like to find the global minimum of the Bethe free energy under the weak consistency constraints. To see how far we can get, we start by constructing the Lagrangian. To the Bethe free energy F^{B} we add multipliers $\mu_{\beta\to\alpha}$ for the weak consistency constraints (31) and ζ_α and ζ_β for the normalization constraints:

$$\begin{aligned}
&\mathcal{L}(\{\tilde{p}_\beta, \tilde{p}_\alpha, \mu_{\beta\to\alpha}, \zeta_\alpha, \zeta_\beta\}) \\
&= -\sum_\alpha \int dx_\alpha\, \tilde{p}_\alpha(x_\alpha) \log \Psi_\alpha(x_\alpha) \\
&\quad + \sum_\alpha \int dx_\alpha\, \tilde{p}_\alpha(x_\alpha) \log \tilde{p}_\alpha(x_\alpha) - \sum_\beta \int dx_\beta\, (n_\beta - 1)\tilde{p}_\beta(x_\beta) \log \tilde{p}_\beta(x_\beta) \\
&\quad - \sum_\beta \sum_{\alpha\in N_\beta} \mu_{\beta\to\alpha}^\top \left[\int dx_\alpha\, \phi_\beta(x_\beta)\tilde{p}_\alpha(x_\alpha) - \int dx_\beta\, \phi_\beta(x_\beta)\tilde{p}_\beta(x_\beta) \right] \\
&\quad - \sum_\beta \zeta_\beta \left[1 - \int dx_\beta\, \tilde{p}_\beta(x_\beta) \right] - \sum_\alpha \zeta_\alpha \left[1 - \int dx_\alpha\, \tilde{p}_\alpha(x_\alpha) \right] . \qquad (32)
\end{aligned}$$

Duality theory then suggests that we should maximize w.r.t. the Lagrange multipliers and minimize w.r.t. the primal variables to find an approximation of $-\log Z$:

$$-\log Z \approx -\log \tilde{Z} = \min_{\{\tilde{p}_\alpha, \tilde{p}_\beta\}} \max_{\{\mu_{\beta\to\alpha}, \zeta_\alpha, \zeta_\beta\}} \mathcal{L}(\{\tilde{p}_\beta, \tilde{p}_\alpha, \mu_{\beta\to\alpha}, \zeta_\alpha, \zeta_\beta\}) .$$

This saddle-point problem is rather difficult to solve since the objective is non-convex in $\{\tilde{p}_\alpha, \tilde{p}_\beta\}$, due to the concave entropy terms for inner clusters. If it were convex, we could exchange the order of min and max, hoping that we could first solve the minimization with respect to $\{\tilde{p}_\alpha, \tilde{p}_\beta\}$ and then maximization with respect to $\{\mu_{\beta\to\alpha}, \zeta_\alpha, \zeta_\beta\}$. But alas, changing the order is not allowed.

Necessary, but not sufficient, conditions for the global minimum of the Bethe free energy under constraints are that the derivatives of the Lagrangian are zero. Therefore, we restrict in this section our analysis to stationary points of the Lagrangian. We will derive fixed point iteration schemes that can be seen as heuristics for finding local minima of the free energies. These fixed point iteration schemes, or message passing algorithms as they are also referred to, are not guaranteed to converge. But if they do, they tend to be a lot faster than the more involved algorithms that are guaranteed to converge. One motivation for the algorithms based on fixed point iteration is that, if the underlying model is a tree and the constraints are strong, they coincide with the algorithm from Section 2 and hence produce exact results in an efficient manner.

4.2.2 Stationary points of the Lagrangian

Necessary conditions for a minimum of the Bethe free energy under the consistency constraints can be found by considering the zero derivative points of the Lagrangian (32). Setting the derivative

$$\frac{\partial \mathcal{L}}{\partial \tilde{p}_\alpha(x_\alpha)} = -\log \Psi_\alpha(x_\alpha) + \log \tilde{p}_\alpha(x_\alpha) + 1 - \sum_{\beta\in N_\alpha} \left[\mu_{\beta\to\alpha}^\top \phi_\beta(x_\beta) \right] - \zeta_\alpha$$

to 0, and replacing ζ_α by its maximum (which implies the normalization of $\tilde{p}_\alpha$) gives

$$\begin{aligned}
\tilde{p}^*_\alpha(x_\alpha; \{\mu_{\beta\to\alpha}\}) &= \frac{1}{Z_\alpha(\{\mu_{\beta\to\alpha}\})} \Psi_\alpha(x_\alpha) \prod_{\beta\in N_\alpha} \mathrm{e}^{\mu_{\beta\to\alpha}^\top \phi_\beta(x_\beta)}, \quad \text{with} \qquad (33) \\
Z_\alpha(\{\mu_{\beta\to\alpha}\}) &= \int dx_\alpha\, \Psi_\alpha(x_\alpha) \prod_{\beta\in N_\alpha} \mathrm{e}^{\mu_{\beta\to\alpha}^\top \phi_\beta(x_\beta)} . \qquad (34)
\end{aligned}$$

Analogously we get

$$\tilde{p}^*_\beta(x_\beta;\{\mu_{\beta\to\alpha}\}) = \frac{1}{Z_\beta(\{\mu_{\beta\to\alpha}\})}\prod_{\alpha\in N_\beta} \mathrm{e}^{\frac{1}{n_\beta-1}\mu_{\beta\to\alpha}^\top\phi_\beta(x_\beta)} \tag{35}$$

$$Z_\beta(\{\mu_{\beta\to\alpha}\}) = \int dx_\beta \prod_{\alpha\in N_\beta} \mathrm{e}^{\frac{1}{n_\beta-1}\mu_{\beta\to\alpha}^\top\phi_\beta(x_\beta)} \ . \tag{36}$$

In the remainder of this section we will drop the explicit dependence of $\{\mu_{\beta\to\alpha}\}$ in $\tilde{p}^*_\alpha(x_\alpha;\{\mu_{\beta\to\alpha}\})$ and $\tilde{p}^*_\beta(x_\beta;\{\mu_{\beta\to\alpha}\})$.

Plugging (33) and (35) into (32) gives, after straightforward manipulations,

$$\mathcal{L}^*(\{\mu_{\beta\to\alpha}\}) = -\sum_\alpha \log Z_\alpha(\{\mu_{\beta\to\alpha}\}) + \sum_\beta (n_\beta-1)\log Z_\beta(\{\mu_{\beta\to\alpha}\}) \ . \tag{37}$$

Setting the partial derivatives of (37) to 0, we get back the weak consistency constraints (31) and the stationary forms (33) and (35) for $\tilde{p}^*_\alpha(x_\alpha)$ and $\tilde{p}^*_\beta(x_\beta)$. Our remaining task is therefore to find an algorithm that makes the factor marginals weakly consistent with the overlap marginals under the constraints that they are of the form (33) and (35). There are several fixed point schemes possible. For instance we could cycle over all α and update all $\mu_{\beta\to\alpha}$ such that after the update the weak consistency constraints hold between α and all its neighbors. In the following section we derive an update scheme that will be shown to correspond to the EP message passing method described in Section 3.1.3.

Perhaps confusingly, it appears that the solution (35) corresponds to a *maximum* of (32) w.r.t. $\tilde{p}_\beta$ rather than a minimum. This apparent contradiction is resolved when we realize that if the constraints are satisfied (i.e., at a stationary point) $\tilde{p}^*_\beta(x_\beta)$ is fully determined by the neighboring $\tilde{p}^*_\alpha(x_\alpha)$ with which they have to be consistent. In other words, if all constraints are satisfied, the overlap marginals $\tilde{p}_\beta$ are functionally dependent on the factor marginals $\tilde{p}_\alpha$, leaving no freedom for maximization nor minimization.

4.3 Equivalence with expectation propagation

From (33) and (10) we have that the Lagrange multipliers $\mu_{\beta\to\alpha}$ are identical to the canonical parameters of the messages that are sent from overlaps to outer clusters in the sum product algorithm:

$$M_{\beta\to\alpha}(x_\beta) = \mathrm{e}^{\mu_{\alpha\to\beta}^\top\phi_\beta(x_\beta)} \ .$$

This motivates the notation for the multipliers.

Suggested by the sum-product framework we can make a change of variables and introduce $\mu_{\alpha\to\beta}$, the messages that are send from outer clusters α to overlaps β. The definition of these messages follows from

$$\mu_{\beta\to\alpha} \equiv \sum_{\alpha'\in N_\beta\setminus\alpha} \mu_{\alpha'\to\beta} \ .$$

That is, the message that β sends to outer cluster α is the product of the messages that β receives from all other outer clusters α'.

Using this substitution we can rewrite (35) as

$$\begin{aligned}\tilde{p}^*_\beta(x_\beta;\{\mu_{\alpha\to\beta}\}) &= \frac{1}{Z_\beta(\{\mu_{\alpha\to\beta}\})}\exp\left[\frac{1}{n_\beta-1}\sum_{\alpha\in N_\beta}\sum_{\alpha'\in N_\beta\setminus\alpha}\mu_{\alpha'\to\beta}\phi_\beta(x_\beta)\right] \\ &= \frac{1}{Z_\beta(\{\mu_{\alpha\to\beta}\})}\exp\left[\frac{1}{n_\beta-1}\sum_{\alpha\in N_\beta}(n_\beta-1)\mu_{\alpha\to\beta}\phi_\beta(x_\beta)\right]\end{aligned}$$

$$= \frac{1}{Z_\beta(\{\mu_{\alpha\to\beta}\})} \exp\left[\sum_{\alpha\in N_\beta} \mu_{\alpha\to\beta}\phi_\beta(x_\beta)\right] \tag{38}$$

$$Z_\beta(\{\mu_{\alpha\to\beta}\}) = \int dx_\beta \, \exp\left[\sum_{\alpha\in N_\beta} \mu_{\alpha\to\beta}\phi_\beta(x_\beta)\right] .$$

We now pick an α and find new outgoing messages $\mu^{\text{new}}_{\alpha\to\beta}$ that make α consistent with its overlaps. In this scheme $\tilde{p}^*_\alpha$ is fully determined by messages that are not changed during this update. To make α weakly consistent with x_β, the new belief over β is set to

$$\tilde{p}^{\text{new}}_\beta(x_\beta) = \mathrm{e}^{\gamma_\beta^\top \phi_\beta(x_\beta)}, \quad \text{with } \gamma_\beta \text{ such that } \langle\phi_\beta(x_\beta)\rangle_{\tilde{p}^{\text{new}}_\beta} = \langle\phi_\beta(x_\beta)\rangle_{\tilde{p}_\alpha} . \tag{39}$$

This process is sometimes referred to as moment matching. The new message from α to β then follows from the above and (38) as

$$\mu^{\text{new}}_{\alpha\to\beta} = \gamma_\beta - \left(\sum_{\alpha'\in N_\beta\setminus\alpha} \mu_{\alpha'\to\beta}\right) . \tag{40}$$

The outer cluster α and overlap β are now weakly consistent, but since $\tilde{p}_\beta$ changed, consistency with other outer clusters will be violated. So updates have to be iterated.

The equivalence between the EP updates, (17) through (18), and the fixed point updates (39) and (40) is immediate. At a fixed point the approximation of $-\log Z$ is given by (37).

The above introduction of the message propagation algorithm in fact follows the arguments of Yedidia et al. (2001) in reverse. In Yedidia et al. (2001) the starting point is a known algorithm (loopy belief propagation) and stationary points of the Bethe free energy are shown to correspond to fixed points of this algorithm. Here we have started with the Bethe free energy and defined a fixed point algorithm such that the correspondence between stationary points and fixed points of the algorithm is by construction.

Expectation propagation in practice often converges, but there is no guarantee that it does. For problems with Bethe free energies with strong consistency constraints, Heskes (2003) shows that stable fixed points of the above algorithm corresponds to local *minima* of the approximate free energy. For problems with weak consistency constraints a similar result is conjectured, but a formal proof is still lacking.

4.4 Equivalence with the expectation consistent approximation

The original EC formulation as discussed in Section 3.2 corresponds to Bethe-type approximation in which we have n factors, each containing the whole domain x. Consequently, there is a single variable $x_\beta = x$ which has all factors as neighbors in the factor graph. Hence $n_\beta = n$. The free energy (37) then boils down to

$$\mathcal{L}^*(\{\mu_{\beta\to\alpha}\}) = -\sum_\alpha \log Z_\alpha(\{\mu_{\beta\to\alpha}\}) + (n-1)\log Z_\beta(\{\mu_{\beta\to\alpha}\}) , \tag{41}$$

with $Z_\beta(\{\mu_{\beta\to\alpha}\})$ from (36). The similarity with the EC free energy from (27) is striking. And indeed, if we substitute the various definitions we find that (27) and (41) are completely equivalent, with Λ_α playing the role of the messages $\mu_{\beta\to\alpha}$.

4.5 Direct minimization of the free energy

The derivation of EP comes with a direct description of an algorithm. However, this algorithm has no guarantee of convergence. EC suggests that we should look for zero derivatives of a functional, the EC free energy (27). By itself it does not tell whether that should be a minimum, maximum,

or saddle-point. The variational approach leads to a specific optimization problem, that we could simply try to solve directly.

A difficulty with directly minimizing (29) is that the objective is in general not convex due to the concave entropy terms

$$-(n_\beta - 1) \int dx_\beta \, \tilde{p}_\beta(x_\beta) \log \tilde{p}_\beta(x_\beta) \, ,$$

that are contributed by the overlaps. Here we introduce the algorithm from Heskes and Zoeter (2002) which is closely related to the CCCP algorithm from Yuille (2002) but makes use of tighter upper bounds.

A slack term $\mathcal{K}$ is used to construct a convex upper bound of $\mathcal{F}$:

$$\begin{aligned} \mathcal{F}^{\text{bound}}(\{\tilde{p}_\alpha(x_\alpha), \tilde{p}_\beta(x_\beta), r_\beta(x_\beta)\}) &= \mathcal{F}(\{\tilde{p}_\alpha(x_\alpha), \tilde{p}_\beta(x_\beta)\}) + \mathcal{K}(\{r_\beta(x_\beta)\}) \\ \mathcal{K}(\{r_\beta(x_\beta)\}) &= \sum_\beta (n_\beta - 1) \sum_{x_\beta} \tilde{p}_\beta(x_\beta) \log \frac{\tilde{p}_\beta(x_\beta)}{r_\beta(x_\beta)} \, . \end{aligned} \tag{42}$$

The weighted KL term $\mathcal{K}$ is guaranteed to be positive if we restrict the r_β's to be proper distributions. Its clever choice effectively cancels the concave parts, resulting in an upper bound *convex* in $\tilde{p}_\alpha$, $\tilde{p}_\beta$, and r_β:

$$\begin{aligned} &\mathcal{F}(\{\tilde{p}_\alpha(x_\alpha), \tilde{p}_\beta(x_\beta)\}) \leq \mathcal{F}^{\text{bound}}(\{\tilde{p}_\alpha(x_\alpha), \tilde{p}_\beta(x_\beta), r_\beta(x_\beta)\}) \\ &= \sum_\alpha \int dx_\alpha \, \tilde{p}_\alpha(x_\alpha) \log \frac{\tilde{p}_\alpha(x_\alpha)}{\Psi_\alpha(x_\alpha)} + \sum_\beta (n_\beta - 1) \sum_{x_\beta} \tilde{p}_\beta(x_\beta) \log r_\beta(x_\beta) \, . \end{aligned}$$

The aim is now to minimize w.r.t. both $\{\tilde{p}_\alpha, \tilde{p}_\beta\}$ and $\{r_\beta\}$, under normalization constraints and weak consistency constraints for $\{\tilde{p}_\alpha, \tilde{p}_\beta\}$. A simple coordinate wise descent does the trick:

Inner loop minimize $\mathcal{F}^{\text{bound}}$ w.r.t. $\{\tilde{p}_\alpha, \tilde{p}_\beta\}$: this is a convex problem with linear constraints which can be solved by any convex minimization procedures. See e.g. Heskes and Zoeter (2002) for some suggestions.

Outer loop minimize $\mathcal{F}^{\text{bound}}$ w.r.t. $r_\beta(x_\beta)$: this is a convex problem. From (42) we see that this minimization step implies a collection of KL minimization problems which is solved by setting $r_\beta(x_\beta) = \tilde{p}_\beta(x_\beta)$ for all β.

4.6 Expectation propagation versus expectation consistency

Since both EP and EC can be derived from a Bethe approximation with weak consistency constraints, we have in fact shown that EP and EC are equivalent. We remind the reader that for the sake of clarity we have restricted the treatment of both methods. EP is not restricted to fully factorizing families, and EC is not restricted to approximations with a single overlap. However, with analogous arguments it should be possible to extend the equivalence to more general cases.

The important difference between EP and EC derives from their motivation. EP is introduced as a procedure for greedily improving localized approximations using projections on tractable distributions. There are many variants of EP that are based on this same idea and do not necessarily have an associated free energy. In some other cases, the free energy functional is derived "after the fact". An example is tree EP (Minka and Qi, 2003), which projects on tree distributions instead of factorized distributions. Welling et al. (2005) shows how tree EP can be derived from free energy functionals. In some cases, as in Zoeter and Heskes (2005), EP-like approximations can only be derived from the energy, not from projection point of view.

EC is motivated by the cavity approach. This cavity interpretation may be useful in a variety of ways. For example, we may be able to argue why the EC approximation gives better results in some applications than for others. E.g., if couplings J_{ij} in an Ising model are fairly short ranged, a central limit argument seems less applicable making the EC approximation less reliable.

Furthermore, the same central limit argument makes it clear why it makes sense to introduce the Gaussian approximation and corresponding factorization that implements it. In fact, there are classes of models for which the cavity approach and thus EC approximation becomes exact in the limit of large N.

5 Discussion and Outlook

The goal of this paper was to review some recent popular and promising approximate inference methods. We wanted to explain the underlying ideas and show how the methods can be unified within a common free energy framework. Such a framework may hopefully stimulate new work in this field, suggesting that many concepts for an approximation can often be extended beyond their original area of application. A unification also satisfies practical needs in machine learning, because it gives us general strategies for developing algorithms for which in some cases convergence can be guaranteed.

However, we also tried to indicate that a general consistent framework for approximations is not everything. It can be applied to a specific problem only after we have chosen appropriate factorizations together with a set of statistics for which consistency is assumed. A justification of the choice of clusters and statistics itself is not a part of such a framework but must come from outside. The obvious requirement of computational tractability cannot be the only guideline. A good amount of intuition about the probabilistic nature of a problem is necessary. The rich experience gained within the area of statistical physics about the behavior of probabilistic models with a large number of variables can be of great help.

Once we have committed ourselves to a specific approximation for a probabilistic model, the accuracy of the method usually remains an open problem. In the literature one often finds empirical studies of good predictive performances of approximate inference algorithms on concrete sets of data. This may not necessarily be attributed to the quality of the underlying approximation. One could think of malicious cases where a bad approximation applied to an insufficient data model would by chance improve the actual prediction on the data set. For sensible applications such as medical expert systems, the computation of a kind of an approximation error or a self-consistent sanity check would be of obvious importance. Possible directions for getting such results could be in the analysis of systematic improvements of approximations, such as higher order Bethe-Kikuchi approximations or perturbative corrections. Alternative approaches could be in the development of statistical, i.e. average case performance measures of approximation methods which would take the random generation of training data into account. Similar to well established concepts in Computational Learning Theory one may think of trying to prove that an approximation is *probably almost accurate*. In the case when statistical errors are large it may often not be necessary to waste computational power on achieving very small approximation errors. Another type of average case analysis could be performed within the Bayesian approach. Here one may e.g. try to show that an approximation achieves an expected loss (over a prior distributions of problem instances) which is close to the Bayes optimal prediction using the correct, but intractable, posterior.

Acknowledgements

The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024. We would like to thank Ruedi Stoop and Bert Kappen for the organization of the PASCAL Workshop on Optimization and Inference in Machine Learning and Physics in Lavin, January 2005, which inspired us to work on the ideas presented in this paper.

References

T. Heskes. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural In-*

formation Processing Systems 15, pages 359–366, Cambridge, 2003. MIT Press. URL `ftp://ftp.snn.kun.nl/pub/snn/pub/reports/Heskes.nips2002.ps.gz`.

T. Heskes, K. Albers, and H.J. Kappen. Approximate inference and constrained optimization. In *Proceedings of Uncertainty in AI*, pages 313–320, 2003.

T. Heskes and O. Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2002)*, San Francisco, CA, 2002. Morgan Kaufmann Publishers.

S. Ikeda, T. Tanaka, and S. Amari. Stochastic reasoning, free energy, and information geometry. *Neural Computation*, 16(9):1779–1810, 2000.

M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

H. J. Kappen and F. B. Rodríguez. Boltzmann machine learning using mean field theory and linear response correction. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 280–286. MIT Press, Cambridge, MA, 1999.

F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

S. L. Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87:1098–1108, 1992.

M. Mézard, G. Parisi, and M. A. Virasoro. *Spin Glass Theory and Beyond*, volume 9 of *Lecture Notes in Physics*. World Scientific, 1987.

T. Minka. The EP energy function and minimization schemes. Technical report, 2001a.

T. Minka and Y. Qi. Tree-structured approximations by expectation propagation. In *Neural Information Processing Systems*, 2003.

T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT Media Lab, 2001b.

K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of Uncertainty in AI*, pages 467–475, 1999.

M. Opper and O. Winther. Adaptive and self-averaging Thouless-Anderson-Palmer mean field theory for probabilistic modeling. *Phys. Rev. E*, 64:056131, 2001a.

M. Opper and O. Winther. Tractable approximations for probabilistic models: The adaptive Thouless-Anderson-Palmer mean field approach. *Phys. Rev. Lett.*, 86:3695, 2001b.

M. Opper and O. Winther. Expectation consistent free energies for approximate inference. In *NIPS 17*, 2005.

G. Parisi and M. Potters. Mean-field equations for spin models with orthogonal interaction matrices. *J. Phys. A: Math. Gen.*, 28:5267, 1995.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan-Kaufmann, 1988.

L.K. Saul, T. Jaakkola, and M.I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.

Y. Teh and M. Welling. The unified propagation and scaling algorithm. In *Advances in Neural Information Processing Systems 14*, page (in press). MIT Press, 2002. URL `http://www.gatsby.ucl.ac.uk/ welling/papers/UPS.ps.gz`.

D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of a 'solvable model of a spin glass'. *Phil. Mag.*, 35:593, 1977.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical report, UC Berkeley, Dept. of Statistics, 2003.

Y. Weiss and W. T. Freeman. Correctness of belief propagation in gaussian graphical models of arbitrary topology. *Neural Computation*, 13(10):2173–2200, 2001.

M. Welling and Y.W. Teh. Belief optimization for binary networks: a stable alternative to loopy belief propagation. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, volume 17, 2001.

M. Welling, Y.W. Teh, and T. Minka. Structured region graphs: Morphing EP into GBP. In *UAI 2005*, 2005.

W. Wiegerinck. Variational approximations between mean field theory and the junction tree algorithm. In *UAI*, 2000.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 689–695, 2001. URL `citeseer.nj.nec.com/yedidia00generalized.html`.

A. Yuille. CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. *Neural Computation*, 14:1691–1722, 2002.

O. Zoeter and T. Heskes. Changepoint problems in linear dynamical systems. Technical report, SNN, Radboud University Nijmegen, 2005.